# Data Science Canvas

**Project:** Forecasting Urban Air Quality for Public Health Advisories

**Team:** Team 14

## Problem Statement

**Business Case & Value Added**
Develop a model to forecast daily AQI 24-72 hours in advance. This enables proactive public health advisories and empowers citizens to reduce exposure.

**Data Landscape**
Required data includes historical air quality metrics and meteorological data (temperature, wind speed). This is available from CPCB monitoring stations

**Model Selection**
1. Baseline Models: Persistence, SMA, EMA, Seasonal Naive
2. Advanced Models: LSTM to capture complex long-term patterns and use weather data as external features.
SARIMA for its ability to handle seasonality.
XGBoost – ensemble model with lag features

**Model Requirements**
The model must accept historical time-series data and output a forecast for the next 1-3 days. It must be validated on an unseen, chronologically later test set.

**Software & Libraries**
Jupyter Notebook, Python, Pandas, NumPy, Matplotlib, Seaborn, Statsmodels, TensorFlow.

**Skills**
Python, Pandas, Scikit-learn, TensorFlow/Keras, Statsmodels, Data Visualization.

## Execution & Evaluation

**Model Evaluation**
Indicators: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE)
Interpretation: Lower values indicate better predictive accuracy.

Category Accuracy : Describes the model AQI prediction that falls into the same category (Higher percentage is recommended)

**Data Storytelling**
Target audience: Public health officials.
Communication: Translate AQI scores into official health categories (e.g., 'Poor', 'Severe') and use clear plots of predicted vs. actual values.

## Data Collection & Preparation

**Data Selection & Cleansing**
Relevant data includes pollutant levels (PM2.5, etc.) and weather data. Data requires cleaning for missing values and aggregation to a daily frequency.

**Data Collection**
Download historical data from CPCB via data.gov.in or Kaggle. Data must be longitudinal (multi-year) to capture seasonal cycles.

**Data Integration**
Data from multiple city stations will be merged and aggregated into a single, chronologically-indexed Pandas DataFrame.

**Explorative Data Analysis**
CPCB AQI calculation
Univariate/Bivariate/Multivariate Analysis
Time-series decomposition
Temporal Analysis
City to city comparison
Impact of COVID-19